



# Tecnológico de Monterrey

## **Análisis de Grandes Volúmenes de Datos**

**Avance de Proyecto 1:**

**Sistema de Recomendación**

**Equipo 39**

**Luis Gerardo Barbosa Mendoza | A01731203**

**Miguel Ángel Marines Olvera | A01705317**

**Genaro Rodríguez Vázquez | A01150931**

**Luis Ángel Seda Marcos | A01795301**

**17 / 05 / 2024**

## 1. Descripción Plan Proyecto

Para llevar a cabo el proyecto donde implementaremos un modelo de recomendaciones, es necesario llevar a cabo ciertos pasos, tales como:

### Recolección y Preparación de Datos

- **Identificar y recopilar datos:** Determinar las fuentes de datos relevantes para el sistema de recomendación, como datos de usuarios, productos, compras, calificaciones, etc.
- **Limpieza y preprocesamiento de datos:** Abordar problemas de calidad de datos como valores faltantes, valores atípicos e inconsistencias.
- **Exploración y análisis de datos:** Visualizar y analizar los datos para comprender las distribuciones, tendencias y relaciones entre las variables.
- **Preparación de datos para el modelado:** Transformar y formatear los datos en un formato adecuado para el modelado de aprendizaje automático.

### Modelado de Recomendaciones

- **Selección de algoritmos de recomendación:** Evaluar y seleccionar algoritmos de recomendación adecuados para el problema, como filtrado basado en contenido, filtrado colaborativo y aprendizaje automático.
- **Entrenamiento y evaluación de modelos:** Entrenar los modelos de recomendación seleccionados en los datos preparados y evaluar su rendimiento utilizando métricas de evaluación relevantes.
- **Selección y optimización del modelo:** Seleccionar el modelo de mejor rendimiento y optimizar sus parámetros para obtener el máximo rendimiento.

### Implementación del Sistema de Recomendación

- **Generación de recomendaciones:** Implementar la lógica para generar recomendaciones personalizadas.
- **Evaluación y mejora del sistema:** Evaluar el rendimiento del sistema de recomendación en un entorno controlado.

### Cronograma del Proyecto:

Las siguientes son las fechas estimadas para la implementación de los pasos del proyecto.

Nota: El cronograma se realizó con lo que se espera se realizará en los siguientes avances de proyecto, sin embargo, el cronograma se actualizará conforme se vaya teniendo acceso a los requerimientos de los siguientes avances de proyecto y a las siguientes actividades para poder planificar de manera adecuada.

TAREAS	MAYO S1	MAYO S2	MAYO S3	MAYO S4	JUNIO S1	JUNIO S2	JUNIO S3	JUNIO S4
Identificar y recopilar datos								
Limpieza y preprocesamiento de datos								
Exploración y análisis de datos								
Preparación de datos para el modelado								
Selección de algoritmos de recomendación								
Entrenamiento y evaluación de modelos								
Selección y optimización del modelo								
Análisis de resultados								
Generación de recomendaciones								
Evaluación y mejora del sistema								

## 2. Descripción Conjunto de Datos

- Descripción del Conjunto de Datos

El conjunto de datos contiene información sobre películas y es útil para desarrollar un programa de recomendación de películas. La información fue obtenida de TMDb y GroupLens y contiene información de películas lanzadas hasta julio del 2017.

El dataset consta de tres archivos en formato CSV. El primer archivo es "credits.csv" e incluye los créditos de las películas y la información de los elencos y equipos de producción. El segundo archivo es "keywords.csv" y proporciona palabras clave relacionadas con las tramas de las películas. El tercer archivo es "movies.csv" y ofrece detalles adicionales sobre cada película.

El dataset se obtuvo de Kaggle, de la siguiente liga:

<https://www.kaggle.com/datasets/rounakbanik/the-movies-dataset?select=keywords.csv>

Nota: El proceso descrito a continuación se puede ver a detalle en la siguiente liga de Google Colab:

<https://colab.research.google.com/drive/1G4J4Vml8wXprquyDm59372mYKmlUN9u1?usp=sharing>

A continuación se describen los archivos del dataset y lo que contienen.

### Credits:

Contiene los créditos de la película y tiene la información del elenco (cast) y equipo de producción (crew).

Contienen 3 columnas o variables:

<b><u>Nombre</u></b>	<b><u>Tipo de Dato</u></b>
Cast (Elenco de la película)	Object
Crew (Equipo de producción de la película)	Object
Movie ID (ID de la película)	Int64

### Keywords:

Contiene palabras clave (keywords) de la trama de la película.

Contienen 2 columnas o variables:

<b><u>Nombre</u></b>	<b><u>Tipo de Dato</u></b>
Keywords (Palabras clave de la trama de la película)	Object
Movie ID (ID de la película)	Int64

### Movies:

Contiene información adicional y detallada de la película.

Contienen 24 columnas o variables:

<b><u>Nombre</u></b>	<b><u>Tipo de Dato</u></b>
Adult (Película para adultos)	Object
Belongs to Collection (Pertenece a una saga la película)	Object
Budget (Presupuesto de la película)	Object
Genres (Géneros de la película)	Object
Homepage (Página web de la película)	Object
Movie ID (ID de la película)	Object
IMDB ID (Identificador único de la película)	Object
Original Language (Idioma original de la película)	Object
Original Title (Título original de la película)	Object
Overview (Sinopsis de la película)	Object
Popularity (Popularidad de la película)	Object

Poster Path (Link a poster de la película)	Object
Production Companies (Compañías productoras de la película)	Object
Production Countries (Países donde se produjo la película)	Object
Release Date (Fecha de estreno de la película)	Object
Revenue (Ganancia económica de la película)	Float64
Runtime (Duración de la película)	Float64
Spoken Languages (Idiomas en la que se encuentra disponible la película)	Object
Status (Estatus de la película cómo estrenada, en producción, etc.)	Object
Tagline (Slogan de la película)	Object
Title (Título de la película)	Object
Video (Archivo de Video)	Object
Vote Average (Calificación promedio que recibe la película)	Float64
Vote Count (Total de calificaciones que recibió la película)	Float64

- Pasos de Preprocesamiento

Nota: El proceso descrito a continuación se puede ver a detalle en la siguiente liga de Google Colab:

<https://colab.research.google.com/drive/1G4J4Vml8wXprquyDm59372mYKmlUN9u1?usp=sharing>

1. Selección de variables relevantes de cada archivo.

<u>Archivo</u>	<u>Columnas o Variables</u>
Keywords	Movie ID, Keywords
Credits	Movie ID, Cast, Crew
Movies	Movie ID, Genres, Original Title, Overview, Production Companies

2. Convertir la variable Movie ID del archivo Movies a Int64 para que esté en el mismo formato que las variables Movie ID de los archivos Keywords y Credits.
3. Creación de un data frame unificado los tres archivos (Credits, Keywords, Movies) con las variables o columnas seleccionadas. La unión se realiza en base a la columna Movie ID de cada archivo.
4. Búsqueda de registros vacíos en las variables.

<u>Variable</u>	<u>Registros Vacíos</u>
Production Companies	4
Overview	995

5. Llenado por conocimiento de los registros vacíos.

Como la variable Production Companies presenta muy pocos registros vacíos, se investigaron los registros faltantes y se llenaron sus registros.

6. Eliminación de un registro vacío

La variable Production Companies presenta un registro vacío, sin embargo, el nombre de la película está en caracteres chinos, por lo que no se entiende el nombre de la película y no se puede hacer la investigación del registro faltante. Como solo es un registro, no hay mayor complicación en eliminarlo.

7. Los valores faltantes de la variable Overview no tienen mayor afectación en el algoritmo de recomendación por lo que se agrega una lista vacía a cada registro, ya que estas películas no tienen público su overview en el idioma en el que se está trabajando (películas sólo publicadas en China).
8. Se convierten a lista los strings de las variables Genres, Keywords y Production Companies para un mejor rendimiento del algoritmo de recomendación.

9. De la variable Crew se obtiene el string del nombre del director y se convierte a lista para un mejor rendimiento del algoritmo de recomendación.
10. De la variable Cast se obtienen los nombres de los actores o personajes (película animada) y se convierten a lista para un mejor rendimiento del algoritmo de recomendación.
11. Las palabras de la variable Overview en formato string se convierten en lista para un mejor rendimiento del algoritmo de recomendación.
12. El conjunto de datos después del preprocesamiento queda de la siguiente manera.

	id	keywords	genres	original_title	overview	production_companies	cast	crew
0	862	[jealousy, toy, boy, friendship, friends, riva...	[Animation, Comedy, Family]	Toy Story	[Led, by, Woody,, Andy's, toys, live, happily,...	[Pixar Animation Studios]	[Tom Hanks, Tim Allen, Don Rickles, Jim Varney...	[John Lasseter]
1	8844	[board game, disappearance, based on children'...	[Adventure, Fantasy, Family]	Jumanji	[When, siblings, Judy, and, Peter, discover, a...	[TriStar Pictures, Teitler Film, Interscope Co...	[Robin Williams, Jonathan Hyde, Kirsten Dunst...	[Joe Johnston]
2	15602	[fishing, best friend, duringcreditsstinger, o...	[Romance, Comedy]	Grumpier Old Men	[A, family, wedding, reignites, the, ancient, ...	[Warner Bros., Lancaster Gate]	[Walter Matthau, Jack Lemmon, Ann-Margret, Sop...	[Howard Deutch]
3	31357	[based on novel, interracial relationship, sin...	[Comedy, Drama, Romance]	Waiting to Exhale	[Cheated, on,, mistreated, and, stepped, on,, ...	[Twentieth Century Fox Film Corporation]	[Whitney Houston, Angela Bassett, Loretta Devi...	[Forest Whitaker]
4	11862	[baby, midlife crisis, confidence, aging, daug...	[Comedy]	Father of the Bride Part II	[Just, when, George, Banks, has, recovered, fr...	[Sandollar Productions, Touchstone Pictures]	[Steve Martin, Diane Keaton, Martin Short, Kim...	[Charles Shyer]



### 3. Exploración del Conjunto de Datos

- Exploración Inicial

Nota: El proceso descrito a continuación se puede ver a detalle en la siguiente liga de Google Colab:

<https://colab.research.google.com/drive/1G4J4Vml8wXprquyDm59372mYKmlUN9u1?usp=sharing>

Dimensiones:

Número de Filas: 4662

Número de Columnas: 8

Variables:

id  
keywords  
genres  
original\_title  
overview  
production\_companies  
cast  
crew

Tipos de Datos:

id (Int64)  
keywords (Object)  
genres (Object)  
original\_title (Object)  
overview (Object)  
production\_companies (Object)  
cast (Object)  
crew (Object)

Porcentaje Datos Faltantes:

id (0%)  
keywords (0%)  
genres (0%)  
original\_title (0%)  
overview (0%)  
production\_companies (0%)  
cast (0%)  
crew (0%)

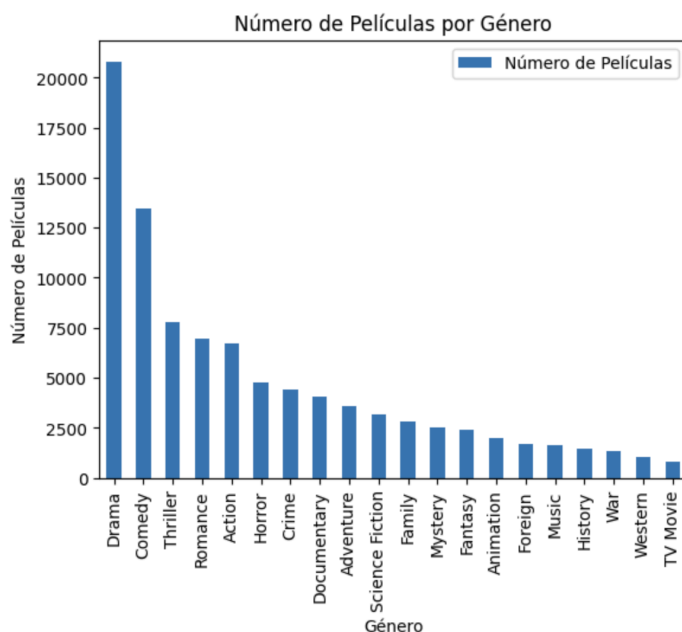
- Análisis del Conjunto de Datos

Nota: El proceso descrito a continuación se puede ver a detalle en la siguiente liga de Google Colab:

<https://colab.research.google.com/drive/1G4J4Vml8wXprquyDm59372mYKmlUN9u1?usp=sharing>

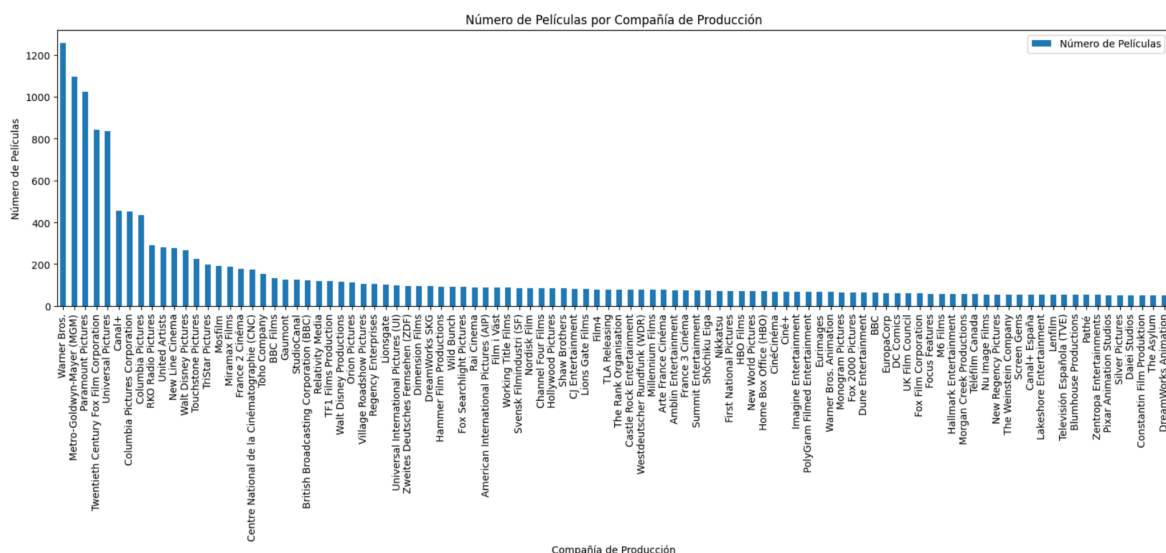
- genres

Son 20 géneros de películas distintos. Se muestra una gráfica del número de películas por género.



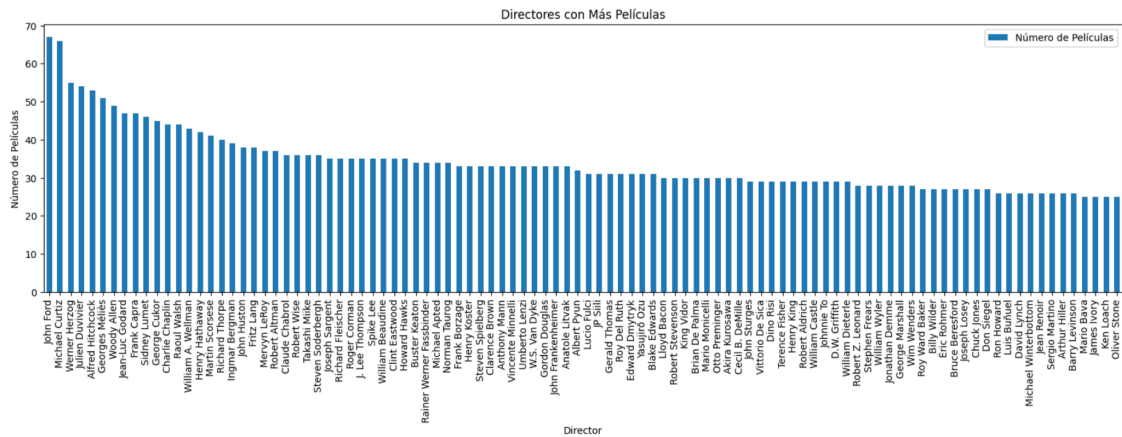
- production\_companies

Hay una variedad bastante grande de compañías de producción de películas. Se muestra una gráfica con las 100 compañías de producción con más películas en orden descendente.



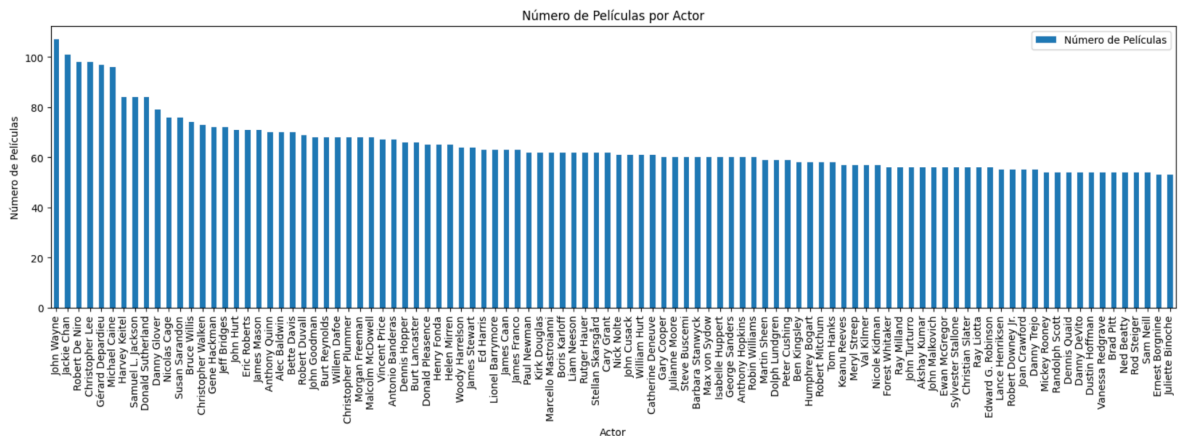
- crew

Hay una variedad bastante grande de directores de películas. Se muestra una gráfica con los 100 directores con más películas en orden descendente.



- cast

Hay una variedad bastante grande de actores y personajes (películas animadas). Se muestra una gráfica con los 100 actores o personajes (películas animadas) con más películas en orden descendente.



- id

Todas las películas tienen un ID distinto, por lo que se tienen 4662 valores diferentes.

- original\_title

Todas las películas tienen títulos distintos, por lo que se tienen 4662 valores diferentes.

- keywords

Son palabras clave de la trama de la película, por lo que se tiene una lista con un vocabulario distinto por película.

- overview

Son palabras de la sinopsis de la película, por lo que se tiene una lista con un vocabulario distinto por película.

## 4. Algoritmo de Recomendación Básico

### Similitud de Coseno

La recomendación de sistema por similitud de coseno es un enfoque utilizado en sistemas de recomendación para calcular la similitud entre dos elementos basándose en sus características. En el contexto de recomendación de películas, por ejemplo, se pueden utilizar las descripciones o metadatos de las películas (como género, actores, director, etc.) como características para calcular la similitud entre ellas.

El cálculo de la similitud de coseno se basa en el concepto de espacio vectorial, donde cada elemento (en este caso, una película) se representa como un vector en un espacio multidimensional, donde cada dimensión corresponde a una característica de la película.

El algoritmo de similitud de coseno mide el ángulo entre dos vectores, lo que refleja la similitud direccional entre ellos. Cuanto más cercano estén los vectores en dirección (es decir, cuanto menor sea el ángulo entre ellos), mayor será la similitud de coseno y, por lo tanto, mayor será la similitud entre los elementos. Para calcular la similitud de coseno entre dos elementos, se utiliza la fórmula:

$$\text{similitud}(\mathbf{A}, \mathbf{B}) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\| \cdot \|\mathbf{B}\|}$$

Donde:

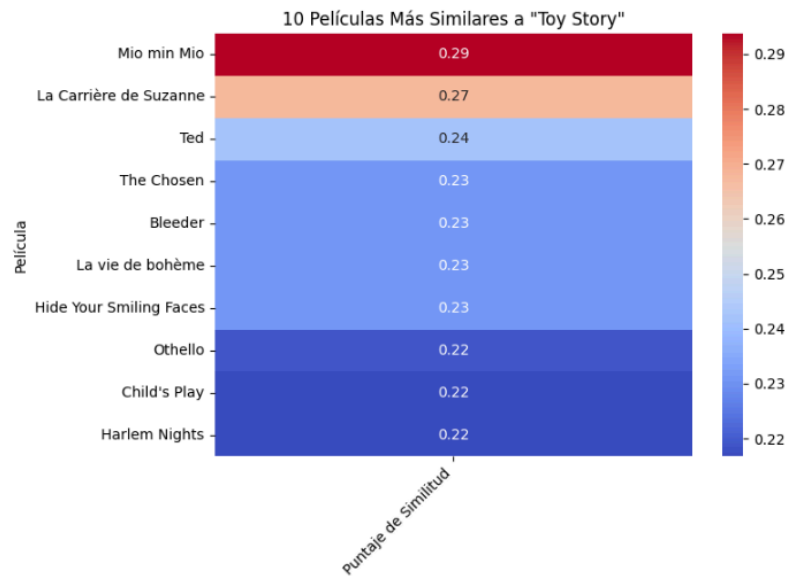
- $\mathbf{A} \cdot \mathbf{B}$  es el producto punto entre los vectores  $\mathbf{A}$  y  $\mathbf{B}$ .
- $\|\mathbf{A}\|$  y  $\|\mathbf{B}\|$  son las magnitudes (normas) de los vectores  $\mathbf{A}$  y  $\mathbf{B}$  respectivamente.

En el contexto de un sistema de recomendación de películas, la similitud de coseno se calcula entre las representaciones vectoriales de las películas (por ejemplo, **utilizando las frecuencias de las palabras clave extraídas de las descripciones de las películas**) para determinar qué tan similares son las películas entre sí. Esto se utiliza para encontrar películas similares a una película de entrada específica y, así, hacer recomendaciones personalizadas.

En términos generales, no hay un valor de similitud de coseno específico considerado "bueno" para aceptar una recomendación. La evaluación de la similitud de coseno depende del contexto del problema y las preferencias del usuario. Se espera que las películas recomendadas tengan una similitud alta entre sí y en comparación con otras películas. Es esencial evaluar tanto los valores numéricos como la relevancia cualitativa de las recomendaciones, considerando las características importantes para el usuario y el contexto del dominio.

Para nuestro ejemplo, utilizaremos la película "Toy Story" como referencia, el algoritmo nos deberá arrojar 10 recomendaciones de películas acorde al título y características de la misma.

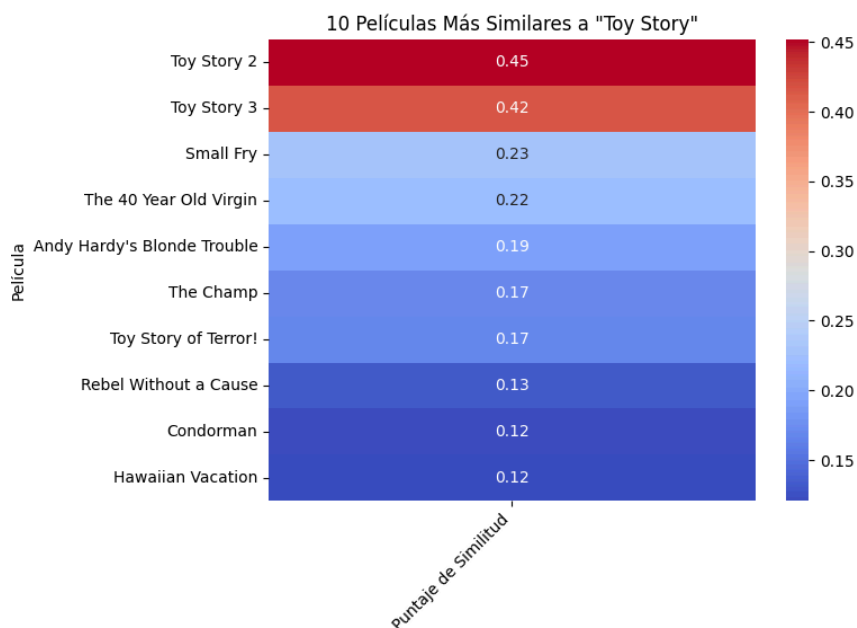
Durante nuestra primera iteración, **al utilizar sólo las palabras clave del dataset**, el resultado no fue el esperado, pues las recomendaciones no nos hacían sentido (falsos positivos):



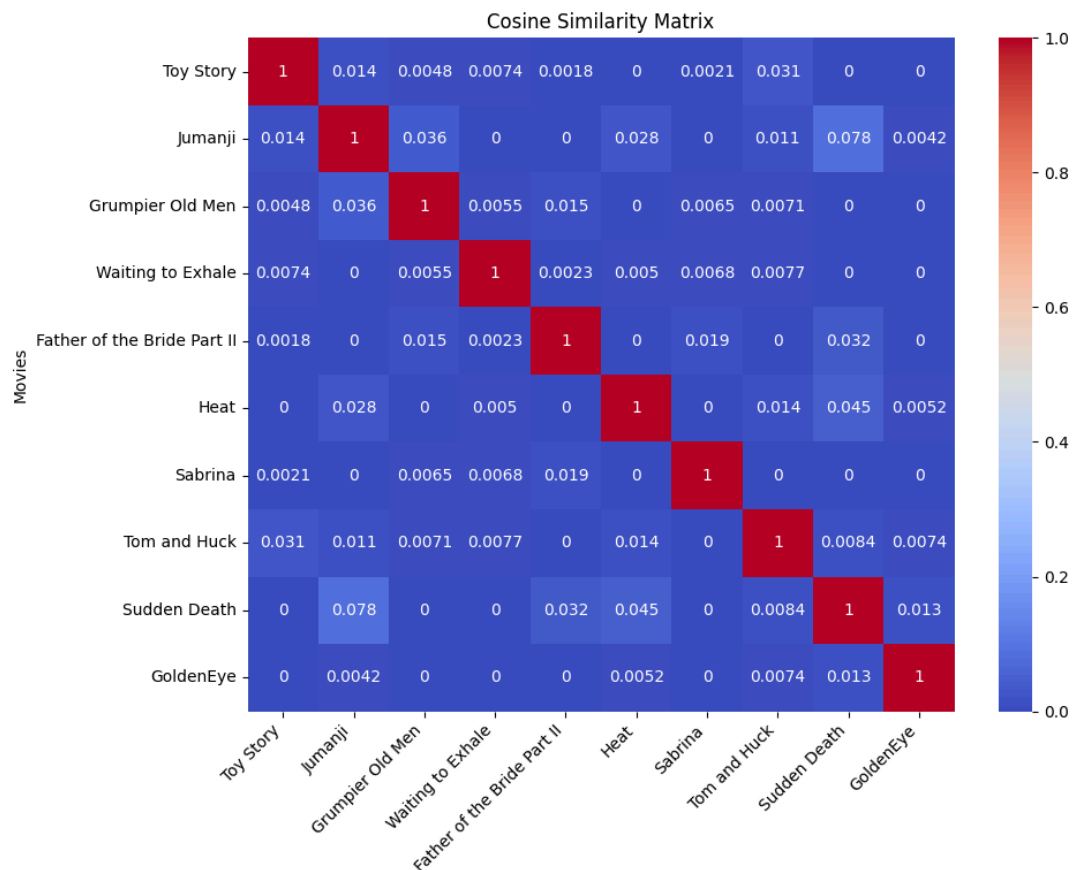
Al analizar el dataset, se llegó a la conclusión de **enriquecer la columna de palabras clave, con otra metadata de la película como el género, la sinopsis, el reparto y el staff.**

```
complete_movies_df['collection'] = complete_movies_df['keywords'] +
complete_movies_df['genres'] + complete_movies_df['overview'] +
complete_movies_df['cast'] + complete_movies_df['crew']
```

El resultado al hacer esta modificación, fue más preciso que durante la iteración anterior:



Como medida comparativa, graficamos también el heatmap para otras 10 películas, para observar qué tan despegados están los coeficientes en películas que no se recomiendan:



**\*El algoritmo se tuvo que recortar a utilizar solamente 25,000 registros debido a la limitante de RAM al utilizar Google Colab.**

Al utilizar el dataset completo (46627 rows × 3 columns) en una PC local con mayores recursos, las recomendaciones cambiaron ligeramente

